Bilingual Word Embedding with Sentence Similarity Constraint for Machine Translation

Kui Wu, Xuancong Wang, AiTi Aw
Institute for Infocomm Research, Singapore
{wuk, wangxc,aaiti}@i2r.a-star.edu.sg

Abstract—**In this work, we propose a context-based bilingual word embedding framework that leverages the information of large amount of parallel sentence pairs which share the same semantic meaning. Such information is abundantly available but has not been fully utilized in previous work of context-based bilingual word embedding models, which only exploit local contextual information through a short window sequence at the word level. To incorporate such information, we define a sentence similarity matching objective which is enforced as a constraint into the original bilingual word embedding objective. They are jointly optimized to better learn the bilingual word embedding. Experimental results show that the proposed model is superior to previous methods on machine translation quality.**

Keywords-word embedding; sentence similarity; machine translation

I. INTRODUCTION

Monolingual word embedding aims to learn a vector representation that captures the meaning of a word from monolingual text [1, 5]. As an extension to it, bilingual word embedding further learns the semantic relationship of word pairs across languages. This property allows one to apply it in cross lingual natural language processing (NLP) tasks such as statistical machine translation (SMT).

In recent years, bilingual word embedding has been applied in SMT and shown to improve the translation quality. A bilingual word embedding method has been proposed by utilizing word alignments to constrain translational equivalence [9]. The learned embedding is incorporated into a phrase-based system by adding a phrase similarity feature into the decoder where the phrase embedding is obtained as an average of the word embedding. A bilingual bag-of-words without alignment (BilBOWA) model has been developed for learning bilingual word embedding which does not require word-aligned parallel training data [2]. It trains directly on monolingual data and extracts a bilingual signal from a smaller set of raw-text sentence-aligned data. It achieves improvement in a lexical translation task. A **c**ontext-**d**ependent bilingual word embedding **m**odel (CDM) which exploits both word alignments and context information has been presented [7]. It is used to initialize a context-dependent convolutional matching (CDCM) model for a phrase-based SMT. Results show that the model outperforms its monolingual counterpart.

In this paper, we present an improved bilingual word embedding framework over the CDM method called **c**ontext-**d**ependent bilingual word embedding **m**odel with **s**entence **s**imilarity constraint (CDM-SS). Under this framework, we exploit the information of the parallel sentence pairs which share the same semantic meaning by incorporating a bilingual sentence similarity constraint on top of the bilingual word embedding objective.

II. METHOD

A. The Context-dependent Bilingual Word Embedding Model (CDM)

The schematic diagram of the CDM architecture is illustrated in Figure 1. Given an aligned word pair (s_i, t_j), the local context information around each word of interest s_i or t_j is first extracted. Let $\overline{s_i} = s_{i-(m-1)/2}, \ldots, s_i, \ldots, s_{i+(m-1)/2}$ and $\overline{t_j} = t_{j-(n-1)/2}, \ldots, t_j, \ldots, t_{j+(n-1)/2}$ be the *m*-word source context window centered at s_i and *n*-word target window centered at t_j. As each word is associated with an index in the respective source or target vocabulary, the first layer of the model maps the context word indices into a feature vector through an embedding lookup operation and concatenating each word embedding to form the vector representation for $\overline{s_i}$ and $\overline{t_j}$, respectively. The meaning of the source input $\overline{s_i}$ and target input $\overline{t_j}$ is then summarized through transformations in their respective semantic embedding space:

$$\mathbf{v}_{\overline{s_i}} = \phi_s(\mathbf{w}_s \cdot g(\overline{s_i})) + \mathbf{b}_s). \quad (1)$$

$$\mathbf{v}_{\overline{t_j}} = \phi_t(\mathbf{w}_t \cdot g(\overline{t_j})) + \mathbf{b}_t) \quad (2)$$

where $g(\cdot)$ is the operation that retrieves the embedding for each input word index and concatenating the embedding

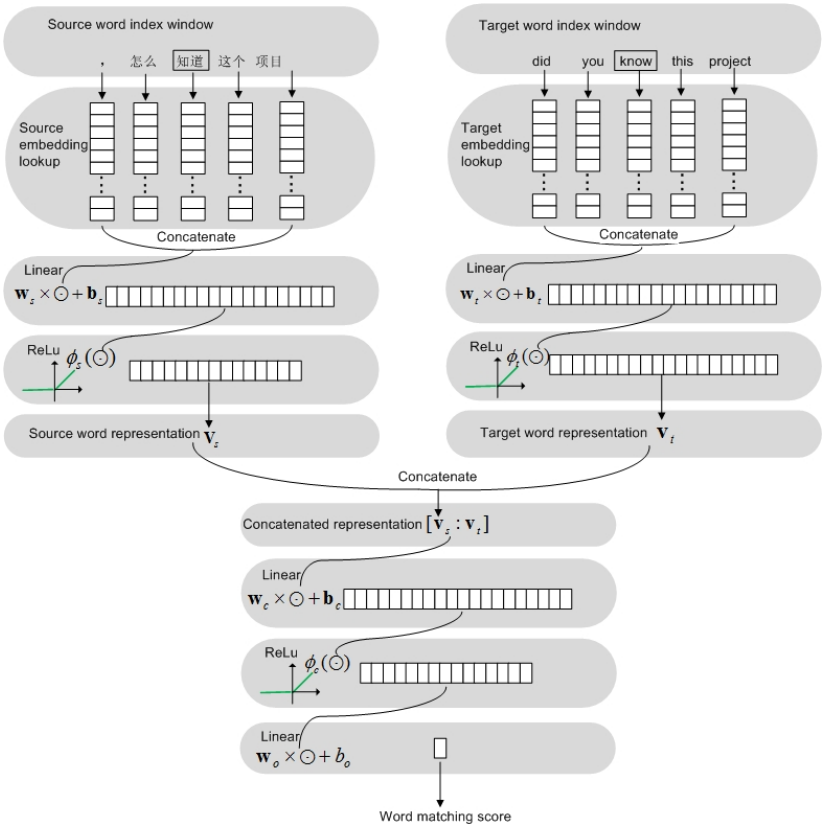

Figure 1. Architecture of the CDM model.

lookup outputs. $\phi_s(\cdot)$ and $\phi_t(\cdot)$ are non-linear activation functions for source and target sequences, respectively. In this work, rectified linear units (ReLU) are used as the activation functions. As compared to sigmoid or tanh functions, ReLU allow for faster and effective training. $\mathbf{w}_s$ and $\mathbf{w}_t$ are the first layer weights of the network for source and target side, respectively. $\mathbf{b}_s$ and $\mathbf{b}_t$ are the biases. The matching score between $\overline{s}_i$ and $\overline{t}_j$ is then measured using a multi-layer perceptron (MLP), which is a nonlinear function for similarity matching. It first combines the feature vectors $\mathbf{v}_{\overline{s}_i}$ and $\mathbf{v}_{\overline{t}_j}$ to get a hidden state:

$$\mathbf{h}_c = \phi_c(\mathbf{w}_c \cdot [\mathbf{v}_{\overline{s}_i} : \mathbf{v}_{\overline{t}_j}]) + \mathbf{b}_c) \quad (3)$$

where ϕ_c is an activation function where ReLU is chosen in this work, $\mathbf{w}_c$ and $\mathbf{b}_c$ are the first layer weight and bias of the MLP. The matching score is then obtained as:

$$f(s,t) = \mathbf{w}_o \mathbf{h}_c + b_o \quad (4)$$

where $\mathbf{w}_o$ and b_o are the output layer weights and bias.

The model is expected to compute a higher matching score when given an aligned word pair (s, t) in a specific context than when given an un-aligned word pair (s,t^*) in the same context. Here, a ranking-based loss is used as the objective function:

$$L(s,t,t^*;\Theta) = \sum_{s,t\in(D_s,D_t)} \sum_{t^*\in V_T} \max(0,\ 1 - f(s,t;\Theta) + f(s,t^*;\Theta)) \quad (5)$$

where $(D_s,\ D_t)$ is the set of all possible aligned word pairs in the specific context from the parallel training corpus. t^* denotes the target context window obtained by replacing the central target word t with a word randomly chosen from the target vocabulary V_T. Θ includes all the parameters of the model: weights and biases of the whole network on both source and target sides and the source and target embedding matrixes. f is the matching score function as defined in (4). The model is trained by minimizing the above objective function.

B. The Proposed Context-Dependent Bilingual Word Embedding Model with Sentence Similarity Constraint (CDM-SS)

The CDM model exploited local contextual information through a short window sequence at the word level. However, we note that the information of large amount of parallel sentence pairs which share the same semantic meaning is not fully utilized in the CDM model. Therefore, we propose to leverage this information to jointly learn the bilingual word embedding. We define a sentence similarity matching score which is enforced as a constraint into the CDM objective. The sentence similarity matching score is

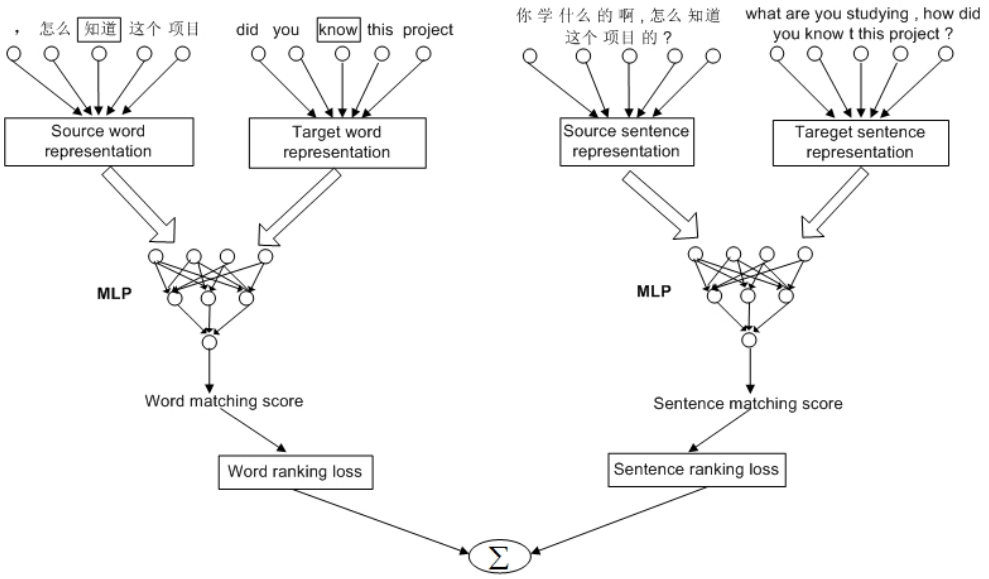

Figure 2. Schematic diagram of the proposed CDM-SS model.

computed in a similar way as the word matching score. As sentences have variable lengths, we take the mean of the individual embedding of each word in a sentence to capture the meaning of the sentence. It is sometimes referred to as vector BOW approach [3]. Given the parallel corpus, our goal is to assign higher matching score for the parallel sentence pair (S, T) while allocating a lower score for an un-parallel sentence (S,T^*). This corresponds to the training objective similar to the CDM model:

$$G(S,T,T^*;\Omega)=\sum_{S,T\in(D_S,D_T)}\sum_{T^*\in D_T}\max(0,\ 1-f(S,T;\Omega)+f(S,T^*;\Omega)) \quad (6)$$

where $(D_S,\ D_T)$ is the set of all parallel sentence pairs in the training corpus. T^* denotes a target sentence randomly chosen from the target sentence corpus D_T. Ω includes all the parameters of the model: weights and biases of the whole network on both source and target sides and the source and target word embedding matrixes. It is noted that the source and target word embedding matrixes are shared among the CDM model and bilingual sentence matching model.

By imposing the bilingual sentence matching score as a constraint to the CDM model, we propose a new architecture called **c**ontext-**d**ependent bilingual word embedding **m**odel with **s**entence **s**imilarity constraint (CDM-SS). The schematic diagram of the proposed CDM-SS model is illustrated in Figure 2. It utilizes both the local context information for meaning-equivalent word pairs and also meaning-equivalent sentence pairs to better learn the bilingual word embedding. The bilingual word embedding objective $L(\cdot)$ and the bilingual sentence matching objective $G(\cdot)$ are jointly optimized:

$$J(s,t,S,T;\Theta,\Omega)=L(s,t,t^*;\Theta)+\lambda G(S,T,T^*;\Omega) \quad (7)$$

III. Experimental Results

A. Setup

We carry experiments on the data obtained from NIST OpenMT15. It is a bilingual Chinese-to-English SMS-Chat social media corpus. The training, tuning, and testing set consist of 130,132, 6,214, and 641 Chinese-English sentence pairs, respectively. The bilingual aligned word pairs are obtained from GIZA++ [6].

For training CDM, input source and target word index window sizes are both set to 5. The input source and target vocabulary contain 16,000 source words and 16,000 target words, respectively. Each source and target word is mapped to a 50-dimensional vector. The first hidden layer for both source and target side consists of 128 neurons. The hidden layer in MLP is 100 dimensional. The source and target word embeddings are initialized using the toolkit Word2Vec [5]. Stochastic gradient descent (SGD) is used to update model parameters after a mini-batch of 50 training sentences with a learning of 0.001. At each epoch, the ranking loss on the tuning set is computed. If the loss is larger than that in the previous epoch, the learning rate is multiplied by 0.5. The model is trained for 20 epochs.

As CDM is a subnet in CDM-SS, the parameter settings of the CDM subset in CDM-SS are the same as described above. For the bilingual sentence matching subset in CDM-SS, the first hidden layer for both source and target side consists of 64 neurons. The hidden layer in MLP is 50 dimensional. Similarly, SGD is used to update the parameters with a mini-batch size of 1000. The learning rate is 0.001. To train CDM-SS, we adopt a sequential training strategy. At each epoch, the subset of CDM is trained first, followed by the bilingual sentence matching subset training. Both subsets access the shared source and

target embedding parameters and update them accordingly. Training is run for 20 epochs.

B. Results on Machine Translation k-best Rescoring

We investigate the influence of the proposed CDM-SS model used in 100-best rescoring in machine translation (MT) as compared to CDM. The experiments are performed on the same data as described in previous subsection which are used to train the CDM and CDM-SS model. The MT system used is Moses, a phrase-based statistical machine translation toolkit [4]. The baseline features we use in decoding include translation models, word and phrase penalty, a distortion model, a lexical reordering model and a 5-gram language model. The MT performance is evaluated on the testing set.

During rescoring, for each pair of source sentence and the corresponding target hypothesis in the 100-best list, the average word matching score for all the aligned source and target words in the sentence pair is used as a rescoring feature for CDM. In addition, a sentence matching score will be extracted for each sentence pair for CDM-SS as another rescoring feature. We re-adjust the weights of the language model feature together with CDM or CDM-SS rescoring features. Other decoding feature weights are kept unchanged, which are the same as in the last round of decoding. The feature weights are tuned by running Z-MERT [8]. The performance comparison is listed in TABLE I. We can see that the CDM rescoring feature does not improve the MT performance over the baseline systems. Instead, the CDM-SS rescoring feature offers consistent gains over the baseline systems with and without rescoring in all metrics. In particular, it achieves an increase of 0. 64 BLEU and 0.43 BLEU over the baseline systems without rescoring and with rescoring, respectively. This shows the superiority of the CDM-SS method over the CDM method by introducing sentence translation equivalence constraint.

TABLE I. EFFECTS OF CDM AND CDM-SS FOR MT 100-BEST RESCORING

System	BLEU	TER
Without rescoring	20.11	67.75
Rescoring with decoding features only	20.32	68.39
Rescoring with additional CDM feature	20.29	68.16
Rescoring with additional CDM-SS feature	20.75	67.71

IV. CONCLUSION

In this work, we have explored the use of sentence translation equivalents as a constraint to guide the learning of bilingual word embedding. We further integrate the bilingual word embedding model into phrase-based SMT system for rescoring. Experimental results show that the proposed model outperforms the model without using sentence similarity constraint. In the future, we would like to apply our model to other tasks in machine translation such as word alignment and phrase table construction.

REFERENCES

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *The Journal of Machine Learning Research,* 3(Feb): 1137-1155, 2003.

[2] S. Gouws, Y. Bengio, and G. Corrado. "BilBOWA: fast bilingual distributed representations without word alignments," In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[3] T. Kenter and M. D. Rijke, "Short text similarity with word embeddings," In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.

[4] P. Koehn, H. Hoang, A. Birch, C. C. Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al., "Moses: Open source toolkit for statistical machine translation," In *Proceedings of ACL (Interactive Poster and Demonstration Sessions*), 2007.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," In *Proceedings of Workshop at ICLR*, 2013.

[6] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, 2003.

[7] B. T. Hu, Z. P. Tu, Z. D. Lu, H. Li, and Q. C. Chen, "Context-dependent translation selection using convolutional neural network," In *Proceedings of ACL-IJCNLP, 536-541, 2015.*

[8] O. F. Zaidan, "Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems," In *The Prague Bulletin of Mathematical Linguistics*, No. 91:79-88, 2009.

[9] W.Y. Zou, R. Socher, D. M. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," In *Proceedings of EMNLP*, pp. 1393-1398, 2013.